SCOPE: Preserving Modality-Specific Cues to Mitigate Modality Laziness in Multimodal Learning

Jingfan Yang[1]* **Rui Zhang**[1]* **Liang Hong**[2]† **Ke Yuan**[1]
[1]School of Information Management, Wuhan University, Wuhan, China
[2]School of Artificial Intelligence, Wuhan University, Wuhan, China
{jingfan_yang, hong, keyuan}@whu.edu.cn, ruizhang8633@gmail.com

Abstract

Multimodal learning aims to learn unified multimodal representations from heterogeneous modalities and supports many natural language processing tasks. However, multimodal models often exhibit *modality laziness*: over-relying on a dominant modality and under-exploiting complementary signals. Existing approaches typically strengthen unimodal training or rebalance modality contributions, but they may still emphasize shared semantics and overlook modality-specific cues. To address this, we propose SCOPE, a unified framework for learning complete multimodal representations, achieving Shared-and-COmplementary cue PrEservation. Firstly, SCOPE uses a mutual information-guided disentanglement module to separate shared semantics from modality-specific cues and mitigate representation collapse. Secondly, SCOPE aligns modalities by enforcing structural consistency between modality-wise semantic graphs, avoiding brittle point-wise matching. Finally, SCOPE performs balanced fusion via structure-aware diffusion attention to integrate shared and complementary cues without feature homogenization. Experiments on five benchmark datasets show that SCOPE consistently outperforms SOTA baselines, achieving up to 27.10% accuracy improvement.

1 Introduction

Multimodal learning aims to build models that jointly understand and reason over heterogeneous modalities, such as text, image, audio, and video (Yuan et al., 2025; Li et al., 2025a; Zhou et al., 2026). It has become a central paradigm for many natural language processing tasks, including sentiment analysis (Zhu et al., 2025; Guo et al., 2025), image captioning (Lee et al., 2024; Yu et al., 2025), and visual question answering (Fan et al., 2024; Mo and Liu, 2024; Li et al., 2025b).

Method	Text-Only	Visual-Only	Multimodal
Concat	74.49	54.14	76.49
AR	74.34	57.23	78.84
DC	74.88	44.51	77.46
Ours	**76.53**	**76.88**	**83.82**

Table 1: Multimodal sentiment analysis accuracy (%) on MVSA. Concat: feature concatenation fusion; AR: architectural regularization; DC: dynamic coordination.

Modality Laziness. Recent studies show that multimodal models may over-rely on a dominant modality whose cues are easier to exploit under the training objective (Huang et al., 2022; Du et al., 2023), leaving other modalities insufficiently leveraged (Zhang et al., 2024b). Although multimodal fusion can improve over a single modality, the non-dominant modality may contribute less than expected. As shown in Table 1, Concat achieves 76.49% in multimodal accuracy, while its visual-only branch is much lower at 54.14%, suggesting that visual cues are not fully translated into the final decision. We refer to this behavior as *modality laziness*: the model predominantly optimizes on one modality, weakening cross-modal complementarity and limiting the benefit of multimodal fusion.

Existing Works and Their Limitations. To mitigate modality laziness, prior work mainly increases the influence of non-dominant modalities during training (Wu et al., 2022). Architectural regularization (AR) strengthens unimodal learning by modifying the model architecture (Wei et al., 2024a; Wei and Hu, 2024; Zhang et al., 2024a,b). Dynamic coordination (DC) adjusts optimization dynamics (e.g., modality re-weighting or gradient balancing) to prevent training from being dominated by a single modality (Lin et al., 2024; Yang et al., 2024b; Wei et al., 2024b; Huang et al., 2025). These methods often improve multimodal prediction, but the non-dominant modality can remain

*Equal contribution.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29066–29078

noticeably weaker. As shown in Table 1, AR achieves 78.84% multimodal accuracy, while the visual-only accuracy remains much lower than the text-only one (57.23% vs. 74.34%). This observation is consistent with the fact that real-world modalities share high-level semantics but also contain complementary cues (Liang et al., 2024): when supervision emphasizes cross-modal agreement, non-dominant modalities can be driven toward shared semantics already captured by the dominant modality, and complementary modality-specific information may be less effectively retained. This leads to two limitations: ❶ *Representation collapse*: over-emphasizing easy shared signals can reduce feature diversity and weaken discriminability. ❷ *Modality-specific underuse*: complementary modality-specific cues are not sufficiently preserved in the unified representation. These limitations motivate us to ask the research question:

How can we learn a unified multimodal representation that aligns shared semantics while preserving modality-specific cues, thereby suppressing modality laziness?

Challenges. To address the research question, we face the following challenges.

- ***Disentangling shared and modality-specific information.*** Shared semantics and modality-specific cues are entangled in unimodal embeddings, making it difficult to separate them while keeping representations discriminative.
- ***Aligning modalities without distorting semantics.*** Modalities differ in distribution and granularity, making it brittle to align them by only pulling paired samples close. Alignment should also preserve cross-modal neighborhood relations among samples.
- ***Fusing modalities without suppressing complementary cues.*** Fusion needs to integrate shared semantics while keeping complementary modality-specific cues, rather than smoothing them into a homogeneous representation.

Our Solution. To address these challenges, we propose **SCOPE** (Shared-and-COmplementary cue PrEservation), a unified multimodal framework that learns representations via *modality-specific disentanglement*, *structure-consistent alignment*, and *balanced fusion*. *First,* we use self- and cross-attention to highlight salient signals and introduce a mutual information-guided disentanglement module to separate shared semantics from complementary modality-specific cues, thereby mitigating representation collapse. *Second,* we propose a structure-consistent alignment strategy to align modalities without distorting their intrinsic geometry. Specifically, we construct a semantic graph for each modality and enforce structural consistency across modalities instead of only pair-wise matching. We theoretically derive that the graph construction is parameter-free. *Finally,* we conduct balanced fusion via structure-aware diffusion attention on the aligned structures, which propagates cross-modal information adaptively while avoiding feature homogenization.

We evaluate SCOPE against eight baselines on five widely used multimodal datasets. SCOPE achieves up to 27.10% accuracy improvement, and ablations further validate each component.

We make the following contributions:

- We propose SCOPE, a unified framework for learning multimodal representations beyond shared semantics to suppress modality laziness.
- We develop an MI-guided disentanglement module to preserve complementary modality-specific cues and mitigate representation collapse.
- We introduce a parameter-free structural alignment and diffusion-attention fusion for robust alignment and balanced integration.

2 Related Work

Modality Laziness. Modality laziness refers to the behavior that multimodal models over-rely on a dominant modality, leaving other modalities under-optimized (Zhou et al., 2023b; Zhang et al., 2024c; Park et al., 2025). In extreme cases, multimodal training may even underperform strong unimodal counterparts (Peng et al., 2022; Hwang et al., 2025). To address this problem, many methods have been proposed, which mainly fall into two main directions. ❶ *Architectural regularization* modifies structures to strengthen unimodal learning, e.g., alternating unimodal optimization (Zhang et al., 2024b), adding auxiliary components to refine cross-modal associations (Yang et al., 2024a; Saeed et al., 2022), performing pre-alignment of unimodal embeddings (Hannan et al., 2025), or modifying the architecture of the fusion module (Ma et al., 2022). ❷ *Dynamic coordination* adjusts optimization across modalities, e.g., adjusting modality-wise learning rates (Wang et al., 2020), reliability-based re-weighting (Arevalo et al., 2017; Yang et al., 2024b; Huang et al., 2025), or per-

forming gradient modulation (Peng et al., 2022; Wei et al., 2024b). Despite advances, these mainly improve non-dominant modality participation, but often do not explicitly preserve complementary modality-specific cues in the unified space.

Multimodal Feature Disentanglement. Multimodal representations often mix shared semantics with modality-specific cues (Hazarika et al., 2020). A common goal is to obtain a shared-specific decomposition or reduce cross-modal redundancy. Existing methods often achieve this either by imposing distance-based or similarity-based regularization in the embedding space (Li et al., 2023; Zhou et al., 2023a; Wang et al., 2025), or by regulating cross-modal statistical dependence, where mutual-information-related objectives are widely used (Yang et al., 2025). For example, ConFEDE (Yang et al., 2023) defines similar and dissimilar features by constructing positive and negative examples, while Uni-Code (Xia et al., 2023) optimizes the loss of mutual information by constructing time-series versions of the upper and lower bounds on the use of mutual information approximations. Inspired by the above methods, we adopt an MI-guided objective to disentangle shared and modality-specific cues.

3 Preliminaries

Multimodal Learning. Let $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be a multimodal dataset, where each sample $\mathbf{x}_i = (x_i^1, \ldots, x_i^M)$ contains M modalities and $\mathbf{y}_i$ is the target label. We learn a multimodal predictor $f(\cdot;\Theta)$ (with unimodal encoders $\{f_m(\cdot;\theta_m)\}_{m=1}^{M}$ and a fusion head $f_0(\cdot;\theta_0)$). The goal is to learn the parameters $\Theta = \{\theta_0, \theta_1, \ldots, \theta_M\}$ by minimizing the empirical risk:

$$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f(\mathbf{x}_i;\Theta),\, \mathbf{y}_i)\,, \tag{1}$$

where $\mathcal{L}$ is cross-entropy loss of model training.

Semantic Graph. A semantic graph provides a discrete proxy of the local geometry of an embedding space. Given unimodal embeddings $\{\mathbf{h}_i^m\}_{i=1}^{N}$ for modality m, we define a semantic graph by a non-negative adjacency matrix $\mathbf{G}^m = [g_{ij}^m] \in \mathbb{R}_+^{N\times N}$ with row-stochastic constraints $\sum_j g_{ij}^m = 1$. Let $f_{ij}^m = \|\mathbf{h}_i^m - \mathbf{h}_j^m\|_2^2$ be the pairwise cost. Graph construction is commonly formulated as:

$$\min_{\mathbf{G}^m} \sum_{i,j} g_{ij}^m f_{ij}^m + \sum_{i,j} \psi(g_{ij}^m), \quad \text{s.t.} \sum_j g_{ij}^m = 1,\ g_{ij}^m \geq 0, \tag{2}$$

where $\psi(\cdot)$ is a regularizer to prevent degenerate weight assignment and encourage locality/sparsity. The objective is typically solved row-wise, i.e., optimizing $\{g_{ij}^m\}_j$ for each node i under simplex constraints (Li et al., 2022).

4 Method

Figure 1 overviews our proposed SCOPE, which consists of three modules: *modality-specific disentanglement*, *structure-consistent alignment*, and *balanced fusion*. The first module extracts discriminative unimodal representations via a mutual information-guided objective to suppress collapse. The second aligns modalities by matching neighborhood structure rather than only paired representations. The third performs structure-guided fusion with diffusion attention to integrate complementary cues while preserving shared semantics.

4.1 Modality-Specific Disentanglement

Prior work mainly balances modality contributions, but may not ensure that non-dominant modalities retain complementary cues beyond shared semantics. To suppress modality laziness, we therefore aim to preserve modality-specific information while maintaining cross-modal semantic agreement. A key challenge is that shared and modality-specific factors are entangled in unimodal embeddings. We therefore refine unimodal features and optimize a mutual information (MI)-guided objective that strengthens matched cross-modal dependence and suppresses cross-sample dependence.

Attention-Based Feature Refinement. For modality m, the unimodal encoder outputs $\mathbf{u}^m \in \mathbb{R}^{N\times d}$, where $\mathbf{u}_i^m$ is the representation of sample i. We apply a lightweight refinement module (self-attention followed by cross-attention) to obtain salient features $\mathbf{h}^m \in \mathbb{R}^{N\times d_h}$. The cross-attention module also yields modality weights $\boldsymbol{\rho} = [\rho^1, \ldots, \rho^M]$, which we reuse in fusion to modulate each modality's contribution.

MI-Guided Disentanglement. We regulate statistical dependence between cross-modal representations. For modality pair (m, n), we encourage high dependence between representations of the *same* sample to preserve shared semantics, while suppressing dependence across *different* samples to prevent collapse and retain modality-specific cues. Formally, we consider the objective

$$\max_{\Theta} \sum_{m\neq n} \Big(\mathbb{E}_i\big[I(\mathbf{h}_i^m;\mathbf{h}_i^n)\big] - \mathbb{E}_{i\neq j}\big[I(\mathbf{h}_i^m;\mathbf{h}_j^n)\big]\Big), \tag{3}$$

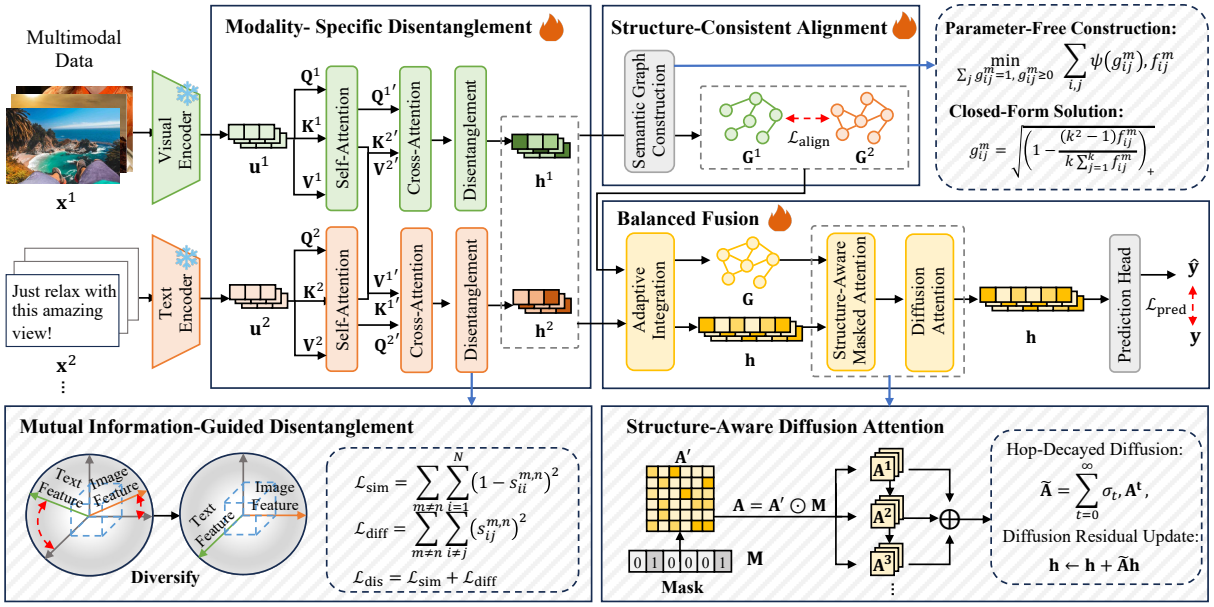

Figure 1: Overview of the proposed SCOPE, which consists of three components: modality-specific disentanglement; structure-consistent alignment; and balanced fusion.

where $I(\cdot;\cdot)$ is the Shannon mutual information (MI). The first term of Eq.(3) promotes cross-modal agreement on shared semantics, and the second term suppresses cross-sample dependence that can lead to collapsed, overly shared representations. Directly optimizing $I(\cdot;\cdot)$ is intractable, so we adopt a similarity-based surrogate.

Similarity-Based Dependence Surrogate. Mutual information measures statistical dependence, but is hard to estimate in high dimensions (Zhou et al., 2025). We therefore use pairwise similarity as a tractable proxy: higher similarity indicates stronger dependence, while near-zero similarity suggests weak dependence. For each modality pair (m, n), let $s_{ij}^{m,n} = \mathrm{sim}(\mathbf{h}_i^m, \mathbf{h}_j^n)$ (cosine similarity by default) and denote $\mathbf{S}^{m,n}$ the similarity matrix. We encourage an identity-like structure of $\mathbf{S}^{m,n}$ by

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{diff}},$$

where $\mathcal{L}_{\text{sim}} = \sum_{m\neq n}\sum_{i=1}^{N}(1 - s_{ii}^{m,n})^2$ strengthens dependence for matched pairs and $\mathcal{L}_{\text{diff}} = \sum_{m\neq n}\sum_{i\neq j}(s_{ij}^{m,n})^2$ suppresses cross-sample dependence. This surrogate preserves the dependence pattern in Eq. (3) and is stable to optimize.

Remark. We adopt a similarity-based orthogonality-style objective as a sample surrogate to implement the same dependence pattern as Eq. (3): high similarity for matched pairs and low similarity for mismatched pairs. We provide the detailed analysis in Appendix A.

4.2 Structure-Consistent Alignment

Cross-modal alignment is necessary for comprehensive representations, but modalities may differ in granularity and exhibit distinct manifold geometries. Thus, aligning only paired representations can misalign semantics by overlooking relational structure. We instead align modalities at the *structure level*: we construct a sparse semantic graph per modality to preserve its local geometry, and then enforce cross-modal structural consistency.

Semantic Graph Construction. When constructing semantic graphs via Eq. (2), the regularizer $\psi(\cdot)$ is introduced to avoid degenerate edge weighting. If ψ is poorly chosen (or $\psi \equiv 0$), the optimization can admit trivial solutions, e.g., concentrating mass on a single neighbor. To avoid this, many graph construction methods introduce hyperparameters (e.g., temperature or kernel bandwidth) to control edge sharpness. However, these hyperparameters are sensitive to embedding scale and modality granularity, and typically require careful dataset- and modality-specific tuning. This tuning burden is undesirable in multimodal settings, where different modalities may exhibit different scales and neighborhood densities. We propose a parameter-free construction by choosing a fixed regularizer $\psi(g) = \log\left(\frac{1+g}{1-g}\right)$. The objective of semantic graph construction is then formulated as:

$$\min_{\sum_j g_{ij}^m=1,\ g_{ij}^m\geq 0} \sum_{i,j} \psi(g_{ij}^m)\, f_{ij}^m. \tag{4}$$

This choice yields a closed-form solution:

$$g_{ij}^m = \sqrt{\Big(1 - \frac{2f_{ij}^m}{\lambda_i^m}\Big)_+}, \tag{5}$$

where $(\cdot)_+ = \max(\cdot, 0)$ and λ_i^m is determined by $\sum_j g_{ij}^m = 1$. We restrict j to the k nearest neighbors of i where k only controls graph sparsity. We provide the derivation details in Appendix B.1.

Theorem 1 (Parameter-Free and Scale-Invariant Weights)**.** *Let $\mathcal{N}_k(i)$ be the k-NN set and $S_i^m \triangleq \sum_{j\in\mathcal{N}_k(i)} f_{ij}^m$. For the row-wise solution in Eq. (5), the Lagrange multiplier λ_i^m satisfies $\frac{2k}{k^2-1} S_i^m \le \lambda_i^m < \frac{2}{k-1} S_i^m$. Choosing $\lambda_i^m = \frac{2k}{k^2-1} S_i^m$ yields $g_{ij}^m = \sqrt{\Big(1 - \frac{(k^2-1)f_{ij}^m}{k\sum_{t\in\mathcal{N}_k(i)} f_{it}^m}\Big)_+}$.*

Proof. Due to the limited space, please see Appendix B.2 for detailed proof.

Theorem 1 shows that our semantic graph construction is free of tunable weighting hyperparameters and scale-invariant. The edge weights depend only on local relative distances: replacing f_{ij}^m by αf_{ij}^m ($\alpha > 0$) leaves g_{ij}^m unchanged for any $\alpha > 0$.

Semantic Graph Alignment. Given $\mathbf{G} = \{\mathbf{G}^1, \ldots, \mathbf{G}^M\}$, we align modalities by minimizing pairwise structural discrepancy:

$$\mathcal{L}_{\text{align}} = \frac{2}{M(M-1)} \sum_{m<n} \frac{1}{N^2} \|\mathbf{G}^m - \mathbf{G}^n\|_F^2. \tag{6}$$

where $\|\cdot\|_F^2$ is the Frobenius norm. $\mathcal{L}_{\text{align}}$ aligns cross-modal neighborhood structures, serving as a structure-level alternative to direct feature matching and reducing manifold distortion.

4.3 Balanced Fusion

After disentanglement and structure-consistent alignment, we obtain discriminative unimodal features $\{\mathbf{h}^m\}$ and structure-aligned graphs $\{\mathbf{G}^m\}$. Existing fusion often relies on dense cross-sample interactions or deep message passing, which can cause over-smoothing and homogenize features, diluting complementary cues and weakening discrimination. Therefore, inspired by previous work (Ning et al., 2025), we propose a balanced fusion module that combines adaptive multimodal integration, graph-masked attention for local aggregation, and diffusion-based multi-hop propagation to capture global context without deep stacking.

Adaptive Integration. We first unify multimodal features and structures so that subsequent fusion operates on a single, modality-aware representation. We reuse the modality weights $\boldsymbol{\rho} = [\rho^1, \ldots, \rho^M]$ learned in the disentanglement module to modulate each modality's contribution:

$$\mathbf{h} = \sum_{m=1}^{M} \rho^m \mathbf{h}^m, \qquad \mathbf{G} = \sum_{m=1}^{M} \rho^m \mathbf{G}^m,$$

where $\mathbf{h} \in \mathbb{R}^{N\times d_c}$ and $\mathbf{G}$ summarizes complementary neighborhood cues across modalities.

Structure-Aware Masked Attention. To prevent dense interactions from washing out modality-specific cues, we restrict attention to semantically meaningful neighbors supported by $\mathbf{G}$. Specifically, we compute self-attention over samples and mask it by the edge set $\mathcal{E}$ of $\mathbf{G}$. Let $\mathbf{Q} = \mathbf{h}\mathbf{W}_Q$ and $\mathbf{K} = \mathbf{h}\mathbf{W}_K$; the dense attention is:

$$\mathbf{A}' = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_c}}\right).$$

We then apply a binary mask $\mathbf{M} \in \{0,1\}^{N\times N}$ where $\mathbf{M}_{ij} = 1$ if $(i,j) \in \mathcal{E}$, and obtain:

$$\mathbf{A} = \text{softmax}\big(\mathbf{M} \odot \mathbf{A}'\big),$$

which aligns attention with the learned semantic structure and performs 1-hop local aggregation.

Diffusion Attention. To incorporate global semantics without stacking many layers (which may over-smooth features), we propagate information through multi-hop neighborhoods with decayed influence. Specifically, $t \in \{0, 1, 2, \ldots\}$ denotes the hop number, and $\mathbf{A}^t$ aggregates information from t-hop neighbors. We diffuse $\mathbf{A}$ via an exponentially weighted power series:

$$\tilde{\mathbf{A}} = \sum_{t=0}^{\infty} \sigma_t \mathbf{A}^t, \qquad \sigma_t = \xi(1-\xi)^t,\ \xi \in (0,1),$$

where $\sum_{t\ge 0} \sigma_t = 1$ and ξ controls the decay rate: larger ξ emphasizes nearer hops. In practice, we approximate the infinite series by truncating it at a finite diffusion step. We update the fused embeddings with a residual connection:

$$\mathbf{h} \leftarrow \mathbf{h} + \tilde{\mathbf{A}}\mathbf{h}.$$

This diffusion emphasizes near neighbors while still capturing multi-hop semantics, improving robustness to over-smoothing. The resulting $\mathbf{h}$ serves as the final multimodal representation.

Remark. $\tilde{\mathbf{A}}$ is row-stochastic and thus defines a valid attention operator. The decayed diffusion preserves feature variation better than deep stacking, mitigating over-smoothing.

4.4 Overall Training Objective

In this paper, we focus on the supervised classification task, and our framework can easily extend to other multimodal tasks by replacing the task loss accordingly. Specifically, for each sample $i \in \{1, \ldots, N\}$, we feed the final fused representation $\mathbf{h}_i$ into a prediction head to output a probability vector $\hat{\mathbf{y}}_i$. We minimize the cross-entropy loss:

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\hat{\mathbf{y}}_i, \mathbf{y}_i),$$

where $\mathbf{y}_i$ is the target label.

Our training includes the disentanglement loss $\mathcal{L}_{\text{dis}}$ and the structure alignment loss $\mathcal{L}_{\text{align}}$ to preserve complementary cues and maintain cross-modal structural consistency.

$$\mathcal{L} = \phi_{\text{dis}}\mathcal{L}_{\text{dis}} + \phi_{\text{align}}\mathcal{L}_{\text{align}} + \phi_{\text{pred}}\mathcal{L}_{\text{pred}}.$$

The weights $\boldsymbol{\phi} = [\phi_{\text{dis}}, \phi_{\text{align}}, \phi_{\text{pred}}]$ are produced by a lightweight MLP $\Phi(\cdot)$ from the current loss values and normalized by a softmax to ensure positivity and $\sum_i \phi_i = 1$.

4.5 Complexity Analysis

We analyze the complexity of each component in SCOPE. Note that all computations are standard mini-batch operations. Let M be the number of modalities, B the batch size, d the feature dimension, k the average number of kept neighbors per node in each graph, and T the diffusion depth. For the modality-specific disentanglement module, the cross-modal similarity cost is $O(M^2B^2d)$. For the structure-consistent alignment module, the semantic graphs are constructed within each mini-batch and sparsified by top-k neighbors. The top-k selection adds $O(MB^2logk)$ computational cost. For the balanced fusion module, the diffusion attention is implemented as sparse top-k propagation with a small depth T. The sparse diffusion costs $O(TkBd)$, where k controls sparsity and M, T are small constants. Overall, SCOPE maintains moderate computational complexity, which is dominated by batch-level similarity matrices and sparse edges, i.e., $O(M^2B^2d + TkBd)$.

5 Experimental Evaluation

We detail the experimental settings in Section 5.1, compare SCOPE with SOTA baselines in Section 5.2, evaluate SCOPE's design in Section 5.3, and perform an ablation study in Section 5.4.

5.1 Experiment Settings

Datasets. We evaluate SCOPE on five widely used multimodal datasets with diverse modality combinations. ❶ *MVSA* (Niu et al., 2016) contains 4,869 visual-text samples with 3 classes. ❷ *TumEmo* (Yang et al., 2020) contains 195,265 visual-text samples with 7 classes. ❸ *CREMA-D* (Cao et al., 2014) contains 7442 visual-audio samples with 7 classes. ❹ *IEMOCAP* (Busso et al., 2008) contains 10,039 visual-audio-text samples with 5 classes. ❺ *MM-IMDb* (Arevalo et al., 2017) is a text-dominant multimodal dataset, which contains 25,959 visual-text movie samples with 23 genre labels for multi-label classification. We adopt a 7:2:1 split for training, validation, and test sets.

Baselines. We compare SCOPE with representative multimodal fusion and modality-laziness mitigation methods, including: ❶ *Conventional fusion* (Zhao et al., 2024): Sum, Concat, and Late Fusion (LF). ❷ *Architectural regularization*: LFM (Yang et al., 2024a) and MLA (Zhang et al., 2024b). ❸ *Dynamic coordination*: CRMT (Yang et al., 2024b), OGM (Wei et al., 2024b), and InfoReg (Huang et al., 2025).

Metrics. We report accuracy and macro-F1 for classification. Accuracy measures overall correctness, while macro-F1 balances precision and recall across classes and is robust to class imbalance. Higher is better for both metrics.

Implementation. All experiments were conducted on a Linux server equipped with an Intel(R) Xeon(R) Gold 5320 CPU @ 2.20GHz, 1.0 TB RAM, and four NVIDIA A40 GPUs, running Ubuntu 22.04. We extract unimodal embeddings using pretrained encoders (CLIP ViT-B/32 (Radford et al., 2021), ImageBind (Girdhar et al., 2023), and Qwen2.5-Omni-3B (Jin Xu, 2025)). We adopt the AdamW optimizer (Loshchilov and Hutter, 2019) (initial learning rate $1e{-}4$ with decay), batch size $B = 60$, for 100 epochs. The self-attention and cross-attention mechanisms used for feature enhancement are configured with 8 attention heads. We set the number of neighbors for semantic graph construction to $k = 30$, corresponding to half of the batch size. We use a diffusion truncation step $T = 3$ for all datasets. All of the reported experimental results are averaged over five runs using the same set of random seeds across all datasets.

Dataset		Sum	Concat	LF	LFM	MLA	CRMT	OGM	InfoReg	SCOPE
MVSA	Text	72.26	74.49	74.34	72.22	74.34	62.99	74.88	75.07	**76.53**
	Visual	47.98	54.14	53.18	51.45	57.23	52.99	44.51	35.26	**76.88**
	Fusion	77.26	76.49	76.84	77.37	78.84	54.19	77.46	77.84	**83.82**
TumEmo	Text	60.39	63.29	62.19	61.88	61.47	61.78	65.03	65.25	**66.72**
	Visual	39.20	38.21	41.28	41.59	43.02	42.74	36.48	36.45	**63.52**
	Fusion	70.92	69.58	74.72	73.25	67.16	68.58	76.11	76.22	**78.10**
CREMA-D	Audio	60.35	57.80	60.89	59.27	58.20	61.56	58.20	52.15	**63.17**
	Visual	37.10	36.83	70.66	68.55	72.33	61.73	42.88	38.17	**72.85**
	Fusion	71.24	71.91	80.38	79.99	79.7	72.85	71.64	69.76	**84.14**
IEMOCAP	Audio	47.32	42.55	52.76	54.92	54.83	53.66	43.72	42.19	**72.90**
	Text	55.37	64.68	65.74	66.85	66.76	53.48	60.89	64.68	**67.04**
	Visual	30.35	30.71	33.19	31.48	57.23	43.27	41.34	42.56	**61.43**
	Fusion	72.62	72.53	74.63	71.54	73.53	55.47	74.71	73.89	**82.57**

Table 2: Comparison of Accuracy(%) across different fusion methods and datasets. "Audio", "Visual", and "Text" represent the performances of the audio, visual, and text features. "Fusion" refers to the results of multimodal fusion. The underline denotes the second-best overall performance, while the bold font denotes the best performance.

Method	MM-IMDb		
	Text	Visual	Fusion
Sum	0.4947	0.2976	0.5892
Concat	0.5095	0.2935	0.5903
LF	0.5837	0.3218	0.6038
LFM	0.5078	0.3308	0.6361
MLA	0.6074	0.3047	0.6174
CRMT	0.5849	0.3029	0.6400
OGM	0.5947	0.3771	0.6288
InfoReg	0.5833	0.3928	0.6309
SCOPE	**0.6250**	**0.4249**	**0.6817**

Table 3: Comparison of macro-F1 across different fusion methods on MM-IMDb dataset. The underline denotes the second-best overall performance, while the bold font denotes the best performance.

Backbone	MVSA		TumEmo	
	w/o	SCOPE	w/o	SCOPE
CLIP	80.27	83.24 (2.97↑)	78.49	79.86 (1.37↑)
ImageBind	79.58	83.82 (4.24↑)	74.12	78.10 (3.98↑)
Qwen2.5-Omni	79.38	82.66 (3.28↑)	80.43	82.33 (1.90↑)

Table 4: Accuracy (%) of different backbones with or without our SCOPE on *MVSA* and *TumEmo*, where ↑ indicates performance improvements.

5.2 Main Results

Below, we compare SCOPE with all the baselines.

SCOPE Demonstrates Superior Task Performance. Table 2 reports unimodal and fused accuracies on four datasets, from which we observe:

❶ *Conventional fusion is prone to modality dominance.* Unimodal performance can be highly imbalanced, indicating limited use of the weaker modality. For example, on TumEmo, text reaches 62.19% while visual is only 41.28% under LF.

❷ *Architectural regularization improves fusion but may still bias toward shared semantics.* Regularization can encourage non-dominant participation, yet complementary cues are not explicitly preserved, so weak modalities can remain under-optimized. For example, on IEMOCAP, LFM achieves 66.85% on text but only 31.48% on visual.

❸ *Dynamic coordination strengthens fusion but can be uneven across modalities.* Re-weighting or gradient balancing can boost fusion, but the optimization may fluctuate across modalities, leading to unstable unimodal performance. For example, on MVSA, InfoReg attains 77.84% fusion accuracy but reduces the visual branch to 35.26%.

❹ *SCOPE achieves the best overall results.* By preserving complementary modality-specific cues and enforcing structure-consistent alignment, SCOPE improves both unimodal robustness and fused accuracy. For example, on IEMOCAP, SCOPE outperforms CRMT by 19.24% on audio, 13.56% on text, 18.16% on visual, and further improves fusion accuracy by 27.10%.

SCOPE Remains Effective Under Strong Modality Dominance. Table 3 reports unimodal and fused performance on the MM-IMDb dataset. We observe that text is more predictive than image on MM-IMDb for all methods. Importantly, our SCOPE achieves the best unimodal and fusion per-

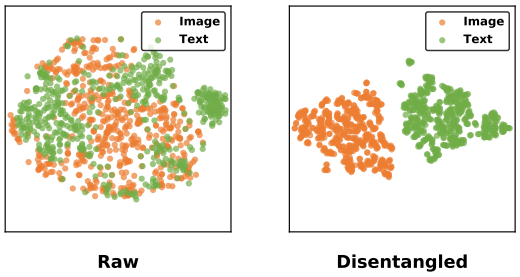

Figure 2: Feature distributions on *MVSA*. The left panel presents the raw embeddings, while the right panel presents the disentangled embeddings.

Dataset	Method	T = 0	T = 1	T = 3	T = 5
MVSA	GAT	5.2259	4.9997	1.2786	0.0314
	DiffuA	5.2259	5.2531	2.4588	1.1961
TumEmo	GAT	10.7721	5.5519	0.5765	0.0499
	DiffuA	10.7721	9.0993	4.5649	2.3025
CREMA-D	GAT	4.2398	2.3846	1.2349	0.1774
	DiffuA	4.2398	3.6404	2.8509	2.4025
IEMOCAP	GAT	3.7057	2.4087	0.5753	0.0064
	DiffuA	3.7057	2.2312	1.4259	1.1924

Table 5: Dirichlet energy of different diffusion steps (T). DiffuA represents the diffusion attention mechanism.

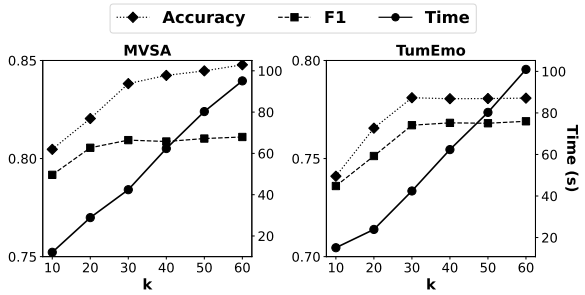

Figure 3: Effect of graph density, where k denotes the number of neighbors used for graph construction, and a smaller k yields a sparser graph.

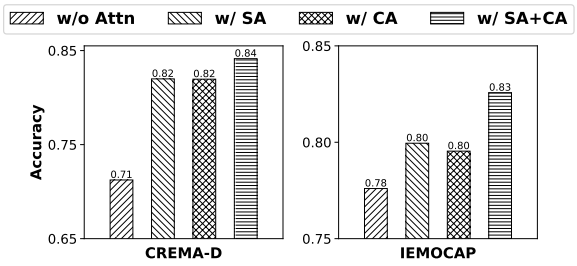

Figure 4: Effect of attention-based feature refinement on *CREMA-D* and *IEMOCAP*, where SA and CA represent self- and cross-attention mechanisms, respectively.

formance among all baselines. For example, compared with the best-performing method (CRMT), SCOPE yields unimodal improvements of 23.57% on average and a fusion improvement of 6.52%.

SCOPE Generalizes Across Backbones. Table 4 reports results of SCOPE with different pretrained encoders on MVSA and TumEmo. SCOPE consistently improves accuracy regardless of the backbone, showing that its gains are not tied to a specific feature extractor. For example, on MVSA, SCOPE improves accuracy by 3.28% with Qwen2.5-Omni and by 4.24% with ImageBind.

5.3 Micro Results

Below, we evaluate the key components of SCOPE.

Effect of Disentanglement. Figure 2 visualizes embeddings before and after applying our mutual information-guided disentanglement in Section 4.1. After disentanglement, embeddings exhibit clearer separation in the representation space. This is consistent with our objective, which strengthens cross-modal agreement for matched samples while suppressing cross-sample dependence, thereby reducing redundant shared signals and retaining modality-specific cues.

Effect of Graph Density. We study how graph sparsity affects structure-consistent alignment by varying the k-NN size used in semantic graph construction. Figure 3 shows a trade-off between accuracy and efficiency: larger k densifies the graph and improves accuracy but increases computational cost. For example, on MVSA, accuracy increases from 80.47% at $k = 10$ to 84.78% at $k = 60$, while the per-epoch time rises from 12s to 95s. We observe diminishing returns: most gains occur when increasing k from 10 to 30, after which improvements saturate. This suggests that a relatively sparse graph already captures the essential neighborhood structure, while overly dense graphs may add redundant or noisy neighbors.

Effect of Diffusion Steps. We examine over-smoothing by measuring the Dirichlet energy of multimodal embeddings produced by our diffusion attention and a graph attention network (GAT)-style stacking baseline while varying the diffusion depth $T \in \{0, 1, 3, 5\}$. Dirichlet energy reflects feature variation over the graph, where higher values indicate less over-smoothing. Table 5 shows that diffusion attention preserves higher energy than GAT, particularly at small T, supporting our default choice of $T = 3$. The formal definition of Dirichlet energy is provided in Appendix C.

Method	MVSA	TumEmo	CREMA-D	IEMOCAP
w/o AFR & MID	77.07	72.24	69.24	77.60
w/o AFR	78.07	74.24	70.16	78.18
w/o MID	83.62	76.06	80.38	81.75
w/o SCA	83.39	75.44	83.16	80.20
w/o BF	82.01	75.04	83.48	82.29
SCOPE	83.82	78.10	84.14	82.57

Table 6: Prediction accuracy (%) on four datasets. ① AFR: attention-based feature refinement; ② MID: MI-guided disentanglement; ③ SCA: structure-consistent alignment; ④ BF: balanced fusion.

5.4 Ablation Study

Below, we evaluate the effect of key modules.

Effect of Attention-Based Feature Refinement. We evaluate the refinement module by enabling self-attention only, cross-attention only, or both, and report the corresponding accuracy on CREMA-D and IEMOCAP. As shown in Figure 4, both self-attention and cross-attention improve performance, and combining them yields the best results. This suggests that the refinement module produces more informative unimodal representations for subsequent multimodal learning.

Effect of Modality-Specific Disentanglement. We evaluate the impact of both attention-based feature refinement (AFR) and MI-guided disentanglement (MID). Table 6 shows that removing either component leads to a clear performance drop across datasets. For example, on CREMA-D, removing AFR reduces accuracy from 84.14% to 70.16%, while removing MID reduces it to 80.38%. Removing both components further degrades accuracy to 69.24%, indicating that AFR and MID are complementary. Overall, AFR enhances salient unimodal information, while MID suppresses redundant dependence and helps preserve complementary cues for downstream fusion.

Effect of Structure-Consistent Alignment. We evaluate the structure-consistent alignment (SCA) module by removing it and replacing the balanced fusion with a standard MLP. As shown in Table 6, accuracy on IEMOCAP drops from 82.57% to 80.20%. This suggests that SCA improves cross-modal correspondence by aligning neighborhood structure, which provides more reliable supervision for subsequent fusion.

Effect of Balanced Fusion (BF). We evaluate the balanced fusion (BF) module by replacing it with mean pooling over node features. As shown in Table 6, this change leads to a clear performance drop. For example, the accuracy drops from 78.10% to 75.04% on TumEmo. This result shows that the balanced fusion improves fusion by integrating structure-guided multi-hop neighbors to capture comprehensive semantics, preventing homogenizing feature distributions.

6 Conclusion

In this paper, we present SCOPE, a unified multimodal learning framework for mitigating modality laziness. Firstly, we design a mutual information-guided disentanglement objective to learn more discriminative unimodal representations. Secondly, we introduce a structure-consistent alignment module that aligns modalities through semantic graphs while preserving local neighborhood geometry. Finally, we propose a balanced fusion module based on structure-aware diffusion attention to effectively integrate shared semantics and complementary cues. Experiments on five widely used multimodal benchmarks against eight representative baselines demonstrate that SCOPE consistently outperforms existing SOTA methods.

Limitations

SCOPE currently assumes aligned multimodal data, where each sample is paired across modalities. Extending it to unaligned or noisy settings is non-trivial, since reliable cross-modal correspondence must be established before disentanglement and structure-consistent alignment. A practical extension is to decouple the pipeline into two stages: correspondence estimation and representation learning. Specifically, pseudo-pairs can first be constructed via cross-modal retrieval in a pretrained embedding space, while noisy matches can be filtered by mutual-consistency checks or similarity thresholding. SCOPE can then be applied to the resulting pseudo-aligned pairs using the same disentanglement, structure-consistent alignment, and fusion objectives. We leave this extension to future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 72474163 and Grant 62276213, the Natural Science Foundation of Hubei Province under Grant JCZRMS202600836, and the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing (TAC)*, 5(4):377–390.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On uni-modal feature learning in supervised multi-modal learning. In *International Conference on Machine Learning (ICML)*, pages 8632–8656. PMLR.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. 2024. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6845–6863.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: one embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. 2025. Bridging the gap for test-time multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 16987–16995.

Abdul Hannan, Muhammad Arslan Manzoor, Shah Nawaz, Muhammad Irzam Liaqat, Markus Schedl, and Mubashir Noman. 2025. Paeff: Precise alignment and enhanced gated feature fusion for face-voice association. *Proc. Interspeech 2025*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 1122–1131.

Chengxiang Huang, Yake Wei, Zequn Yang, and Di Hu. 2025. Adaptive unimodal regulation for balanced multimodal information acquisition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 25854–25863.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably). In *International Conference on Machine Learning (ICML)*, pages 9226–9259. PMLR.

Seong-Hyeon Hwang, Soyoung Choi, and Steven Euijong Whang. 2025. Midas: Misalignment-based data augmentation strategy for imbalanced multimodal learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3732–3746.

YaoChong Li, Yi Qu, Ri-Gui Zhou, and Jing Zhang. 2025a. Qmlsc: A quantum multimodal learning model for sentiment classification. *Information Fusion*, 120:103049.

Yong Li, Yuanzhi Wang, and Zhen Cui. 2023. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6631–6640.

Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. 2025b. Patch-level sounding object tracking for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 5075–5083.

Zhenglai Li, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, and En Zhu. 2022. High-order correlation preserved incomplete multi-view subspace clustering. *IEEE Transactions on Image Processing (TIP)*, 31:2067–2080.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Wenzhong Tang, Zitong Yu, and Alex Kot. 2024. Suppress and rebalance: Towards generalized multimodal face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 211–221.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *The 7th International Conference on Learning Representations (ICLR)*.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186.

Wentao Mo and Yang Liu. 2024. Bridging the gap between 2d and 3d visual question answering: A fusion approach for 3d vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 4261–4268.

Xuying Ning, Dongqi Fu, Tianxin Wei, Wujiang Xu, and Jingrui He. 2025. Graph4mm: Weaving multimodal learning with structural information. *International Conference on Machine Learning (ICML)*.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22 (MMM)*, pages 15–27. Springer.

Simon Park, Abhishek Panigrahi, Yun Cheng, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. 2025. Generalizing from simple to hard visual reasoning: Can we mitigate modality imbalance in vlms? *International Conference on Machine Learning (ICML)*.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8247.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Muhammad Saad Saeed, Muhammad Haris Khan, Shah Nawaz, Muhammad Haroon Yousaf, and Alessio Del Bue. 2022. Fusion and orthogonal projection for improved face-voice association. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7057–7061. IEEE.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. 2025. Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pages 8141–8149.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. 2024a. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27338–27347.

Yake Wei and Di Hu. 2024. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *International Conference on Machine Learning (ICML)*.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. 2024b. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 24043–24055. PMLR.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. 2023. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:63529–63541.

Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7617–7630.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. 2025. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6803–6811.

Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia (TMM)*, 23:4014–4026.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024a. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:62108–62122.

Zequn Yang, Yake Wei, Ce Liang, and Di Hu. 2024b. Quantifying and enhancing multi-modal robustness with modality preference. *The 12th International Conference on Learning Representations (ICLR)*.

Linhao Yu, Xingguang Ji, Yahui Liu, Fanheng Kong, Chenxi Sun, Jingyuan Zhang, Hongzhi Zhang, Fuzheng Zhang, Deyi Xiong, and 1 others. 2025. Evaluating multimodal large language models on video captioning via monte carlo tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6435–6462.

Yuan Yuan, Zhaojian Li, and Bin Zhao. 2025. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34.

Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and Liang Wang. 2024a. Modality-balanced learning for multimedia recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 7551–7560.

Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. 2024b. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27456–27466.

Yedi Zhang, Peter E Latham, and Andrew Saxe. 2024c. Understanding unimodal bias in multimodal deep linear networks. *International Conference on Machine Learning (ICML)*.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36.

Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. 2023a. A unified multimodal de-and re-coupling framework for rgb-d motion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10):11428–11442.

Xiaokai Zhou, Xiao Yan, Fangcheng Fu, Ziwen Fu, Tieyun Qian, Yuanyuan Zhu, Qinbo Zhang, Bin Cui, and Jiawei Jiang. 2025. Ps-mi: Accurate, efficient, and private data valuation in vertical federated learning. *Proceedings of the VLDB Endowment (VLDB)*, 18(10):3559–3572.

Xiaokai Zhou, Xiao Yan, Xinyan Li, Yuxiang Wang, Quanqing Xu, Chuang Hu, Tieyun Qian, and Jiawei Jiang. 2026. HAL: Accurate, private, and efficient sample alignment for multimodal federated learning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM.

Ying Zhou, Xuefeng Liang, Shiquan Zheng, Huijun Xuan, and Takatsune Kumada. 2023b. Adaptive mask co-optimization for modal dependence in multimodal learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. 2025. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 22123–22138.

A MI-Guided Disentanglement

Below we provide additional intuition for mutual information-guided disentanglement.

Mutual information as dependence control. Mutual information quantifies statistical dependence: $I(Z_1; Z_2) = 0$ iff Z_1 and Z_2 are independent. In multimodal learning, paired samples are expected to share semantics and thus exhibit higher cross-modal dependence, whereas mismatched samples should be weakly dependent. This motivates Eq. (3), which increases dependence for matched pairs while suppressing cross-sample dependence.

Why similarity is a practical surrogate. Estimating the mutual information in high-dimensional continuous spaces is typically intractable. We therefore use pairwise similarity as a tractable proxy for dependence: higher similarity generally reflects stronger coupling between representations, while near-zero similarity indicates weak coupling. For modality pair (m, n), we define a similarity matrix $\mathbf{S}^{m,n}$ with $(\mathbf{S}^{m,n})_{ij} = \mathrm{sim}(\mathbf{h}_i^m, \mathbf{h}_j^n)$, and encourage an identity-like structure (large diagonal, small off-diagonal). This enforces the same matched-vs.-mismatched dependence pattern targeted by the MI-guided disentanglement objective in Eq. (3).

Connection to contrastive learning. Contrastive objectives (e.g., InfoNCE) are commonly viewed as variational surrogates for mutual information. Our orthogonality-style loss offers a simpler alternative that avoids explicit negative sampling and directly penalizes off-diagonal similarities, while preserving the same dependence structure.

B Semantic Graph Construction

B.1 Derivation Details of Semantic Graph Construction

By taking the derivative with respect to (w.r.t.) the nonlinear weight term in Eq. (4), we have:

$$\frac{\partial \log\left(\frac{1+g_{ij}^m}{1-g_{ij}^m}\right)}{\partial g_{ij}^m} = \frac{2}{1-(g_{ij}^m)^2} > 0, (0 \le g_{ij}^m < 1),$$

which indicates that weight term $\log\left(\frac{1+g_{ij}^m}{1-g_{ij}^m}\right)$ is proportional to weight g_{ij}^m due to the positive derivative, i.e., monotonically increasing w.r.t. g_{ij}^m. In addition, weight term $\log\left(\frac{1+g_{ij}^m}{1-g_{ij}^m}\right)$ projects the weight $0 \le g_{ij}^m < 1$ to the interval $[\log\left(\frac{1+0}{1-0}\right), \log\left(\frac{1+1}{1-1}\right)) = [0, \infty)$. In

sum, both properties regarding positive proportion and nonnegative value make the nonlinear term $\log\left(\frac{1+g_{ij}^m}{1-g_{ij}^m}\right)$ suitable for a weight learning mechanism.

We further optimize the proposed framework in Eq. (4). The Lagrangian function of Eq. (4) is represented as:

$$\mathbf{L} = \sum_{i,j} \log\left(\frac{1+g_{ij}^m}{1-g_{ij}^m}\right) f_{ij}^m - \sum_i \lambda(\sum_j g_{ij}^m - 1) - \sum_{i,j} \delta_{ij} g_{ij}^m,$$

where λ and δ_{ij} are Lagrangian multipliers. Accordingly, the formulation to update g_{ij}^m could be summarized as:

$$g_{ij}^m = \sqrt{(1 - \frac{2f_{ij}^m}{\lambda})_+},$$

which equals to Eq. (5).

B.2 Proof of Theorem 1

Proof. Since f_{ij}^m is negatively proportional to g_{ij}^m in Eq. (5), i.e., $f_{ij}^m \propto -g_{ij}^m$, we have $g_{i1}^m \geq g_{i2}^m \geq \cdots$ under $f_{i1}^m \leq f_{i2}^m \leq \cdots$. Due to $g_{ij}^m < 1$, the following inequality can be inferred that:

$$\begin{aligned} g_{ij}^m &= \sqrt{(1 - \frac{2f_{ij}^m}{\lambda})_+} \geq (1 - \frac{2f_{ij}^m}{\lambda})_+ \\ \Rightarrow \sum_j (1 - \frac{2f_{ij}^m}{\lambda})_+ &< 1 = \sum_j g_{ij}^m. \end{aligned} \tag{7}$$

By considering only k nearest neighbors, i.e., $\sum_{j=1}^k g_{ij}^m = 1 = \sum_{j=1}^k \sqrt{(1 - \frac{2f_{ij}^m}{\lambda})}$, we have:

$$1 = \sum_{j=1}^k \sqrt{(1 - \frac{2f_{ij}^m}{\lambda})} \leq \sqrt{k \sum_{j=1}^k (1 - \frac{2f_{ij}^m}{\lambda})}. \tag{8}$$

Based on Eqs. (7) and (8), we could further infer that:

$$\begin{aligned} &\sum_{j=1}^k (1 - \frac{2f_{ij}^m}{\lambda}) < 1 \leq \sqrt{k \sum_{j=1}^k (1 - \frac{2f_{ij}^m}{\lambda})} \\ \Rightarrow &\frac{2k \sum_{j=1}^k f_{ij}^m}{k^2 - 1} \leq \lambda < \frac{2 \sum_{j=1}^k f_{ij}^m}{k - 1}, \end{aligned}$$

Therefore, we could specifically choose the lower bound $\lambda = \frac{2k \sum_{j=1}^k f_{ij}^m}{k^2-1}$. Therefore, the closed-form solution of g_{ij}^m can be represented as:

$$g_{ij}^m = \sqrt{(1 - \frac{(k^2-1) f_{ij}^m}{k \sum_{j=1}^k f_{ij}^m})_+}.$$

which completes the proof. □

C Definition of Dirichlet energy

We introduce the Dirichlet energy to quantify the smoothness of node representations.

Definition 1 (Dirichlet Energy)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with node set $\mathcal{V}$ and edge set $\mathcal{E}$. Let $\mathbf{F}^n \in \mathbb{R}^{|\mathcal{V}| \times d}$ denote the node feature matrix at the n-th layer of a GNN. The Dirichlet energy is defined as:*

$$\mathcal{E}(\mathbf{F}^n) = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} \left\| \mathbf{F}_i^n - \mathbf{F}_j^n \right\|_2^2,$$

where $\mathcal{N}_i$ is the neighborhood of node i. Over-smoothing corresponds to representations becoming nearly identical across neighbors, in which case $\mathcal{E}(\mathbf{F}^n) \to 0$.